下载APP

07 | 散点图和相关性：怎样快速从数据当中找到规律？

2021-08-11 郭炜

《数据分析思维课》 课程介绍 >

讲述：郭炜

时长 18:18 大小 16.76M

数据给你一双看透本质的眼睛，这里是《数据分析思维课》，我是郭炜。

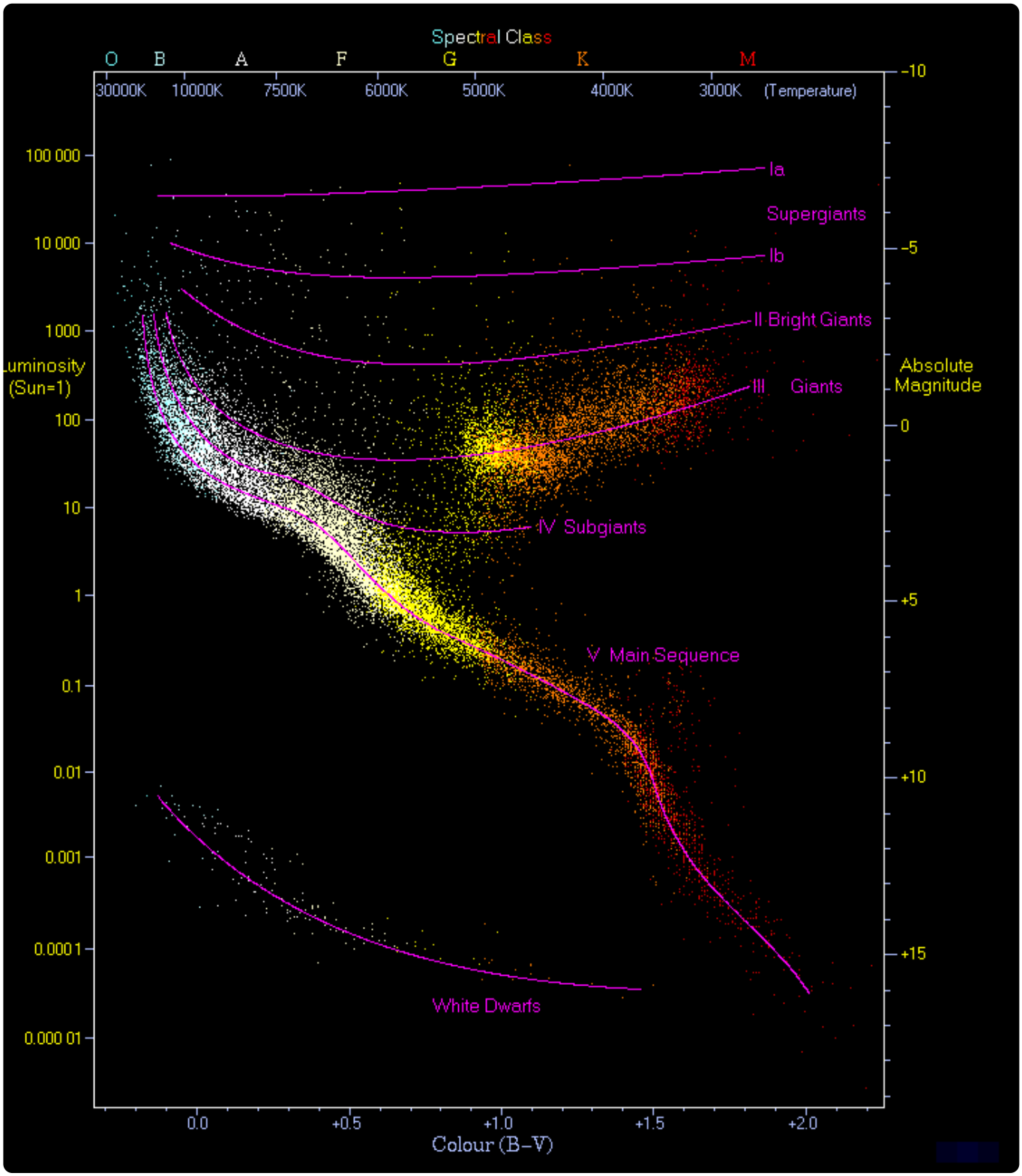
前面我们讲了怎么从一个数据累计量当中，发现它的分布规律。但其实很多时候我们遇到的数据并不是累计的分布数据，而是连续的一些数据，并且我们需要基于这些数据做一些总结和推断，甚至是预测。

比如在工作中，我们要根据成本和收入来预测下半年的投入和产出；在生活里，需要看下自己投资的基金、股票金额和回报的整体关系，又或者看自己体重增长和摄入热量的关系，这些其实都是要从数据当中去找趋势规律。

今天我就来教你一个最简单的发现数据趋势规律的工具，以及这个工具的使用方法——散点图和相关性分析。

散点图的历史

散点图被称之为万图之王。在 1913 年，美国一个叫做亨利·诺利斯·罗素（Henry Norris Russell）的天文学家用散点图把宇宙的趋势给揭示了出来。怎么揭示的呢？罗素同学利用散点图把 2200 颗恒星按光谱和亮度两个参数进行分析，将恒星光度（或绝对星等）为纵轴、以恒星的光谱类型（或表面温度）为横轴，就像下图这个样子。



通过这个散点图，罗素画出了一条趋势线，这条趋势线揭示了恒星从原恒星到红巨星到红白矮星、黑矮星的一个演变的过程，这就是著名的赫罗图。换句话说，这个散点图揭示了恒星这一生的秘密。

看，散点图可以帮我们解释宇宙的秘密。但这还没有结束，后来哈勃（对，就是以他的名字命名哈勃天文望远镜那位），也利用散点图找到了大爆炸理论和解释宇宙膨胀思想的一个关键证据，如果你对这个故事感兴趣，可以去本节课的附录里进一步学习。

散点图的制作原则

散点图能够帮助科学家在这么复杂的宇宙里找到一些客观的规律，更不要说我们日常去预测销售量和成本投入之间的关系、分析自己投资和回报的关系了。

那么我们怎样做出一个正确的散点图呢？其实我们用 Excel 就可以非常方便地去做出散点图。我在后面的实操课程里会进一步手把手地带你去操作，我们今天先重点关注散点图的制作原则。

无论你用什么工具，做散点图有三个最基本的规则。

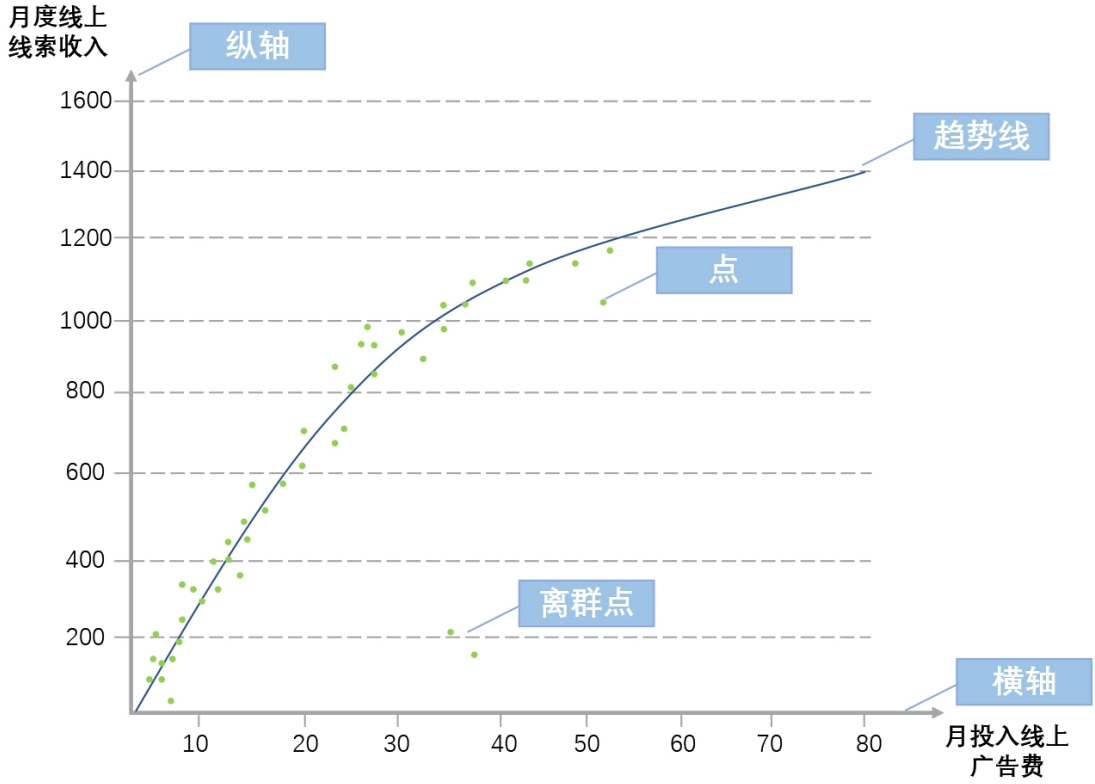
第一，散点图反映的是两个变量之间的关系。因此你要把两个变量分别放在 X 轴和 Y 轴上面，不要有第三个变量放进来进行混淆。当然，散点图的变种——气泡图有更多维度，但是从趋势角度来看，还是要把最重要的两个变量分别放在 X 轴和 Y 轴。

第二，为了能够明确展示数据之间的趋势，我们的 Y 轴必须要从零开始。这点就和很多的柱状图不同了。很多柱状图的坐标轴为了表示数据的清晰，Y 轴可以从一半（比如说 500）开始。**另外，散点图坐标轴颗粒度要合适**，最终聚成一团或者散列太大，都会让我们无法快速找到趋势。

第三，为了表示趋势的清晰，我们一般都会添加一条趋势线来表明背后的规律。说明一个趋势的趋势线只能有一条，不能有多条特别是出现趋势相交的情况。可别小看这个趋势线，这就是画龙点睛的一笔，背后其实是你对业务、数据、算法深刻的理解和认知。画得好，你就是哈勃和罗素，画不好你就成了我接下来会讲到的得克萨斯的伪神枪手。

比如对网站广告投入成本和销售量增加的趋势来说，散点图是要让你看到销售金额随着网站投放成本投入增加逐步增长的情况。它应该有标准的横轴和纵轴，分别代表销售金额和

投放成本投入，有每一个月份对应成本和金额的离散点以及给出的趋势线。可能会有少部分点是离群点，离趋势线有一定的距离，但属于正常情况，如下图所示。

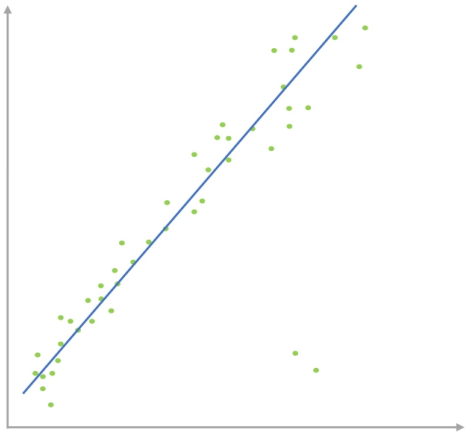


这样我们可以看到何时我们投放网站最有效，而不是所谓的“增长黑洞”：一直广告投入最后回报率很低。

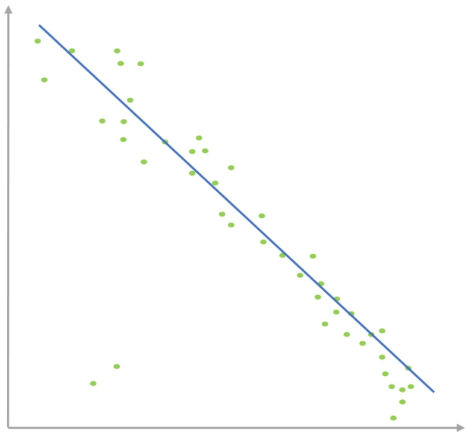
通过散点图寻找规律

有这三个原则的指导，我们画出来的散点图一般不会错。但是光有图肯定不行，我们如何在图中去找到数据之间的关系呢？下面我给你介绍几个散点图里面最常见的数据趋势。

首先是**正相关**，这个你很容易理解，它就是两个系数的变化基本上画出来以后都在一个直线分布上，此增彼涨，两个变量都一起增加，比如我们上下班的距离和时间。同样，**负相关**是一个此增彼减斜着向下的趋势，这种方式也比较容易判断。

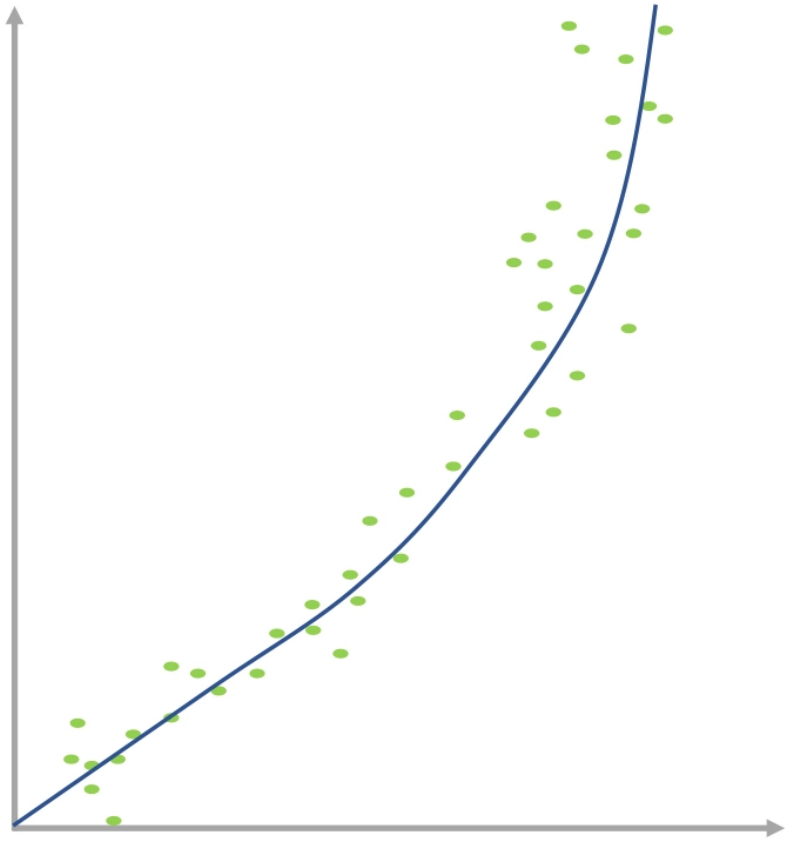


正相关



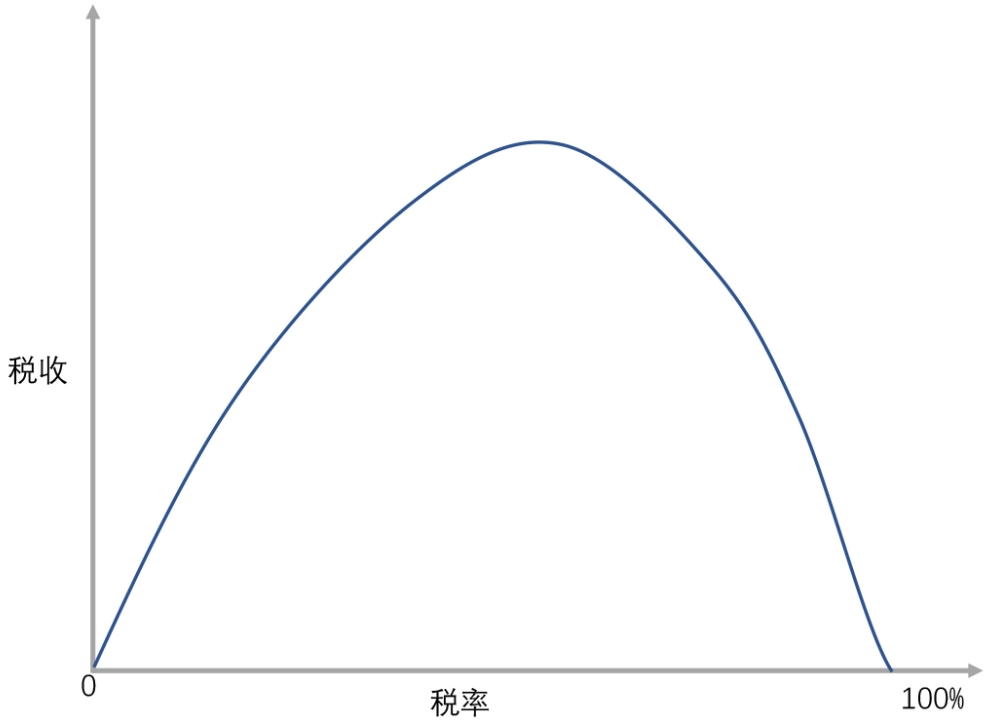
负相关

指数增长和上节课的指数分布有些类似，只不过指数分布计算的是一个加起来的累计分布值，而指数增长指的是两个具体数值的关系。这块如果你还是有些分不清楚的话，可以再复习一下上一节课。你看到这个曲线，一般会很高兴，因为它代表你抓住了一些别人没有的机会。指数级别的变化，一般背后都酝酿着极大的商机和机会。



正 U 型趋势和**反 U 型**趋势，这两个也是比较常见的趋势线。它的样子其实就像字母的 U，或者是倒过来的 U（从零开始从零结束）。一个比较著名的反 U 型趋势就是经济学里面的“拉弗曲线”（ Laffer curve），这条曲线最典型地反映了政府税收收入和税率之间的关系。

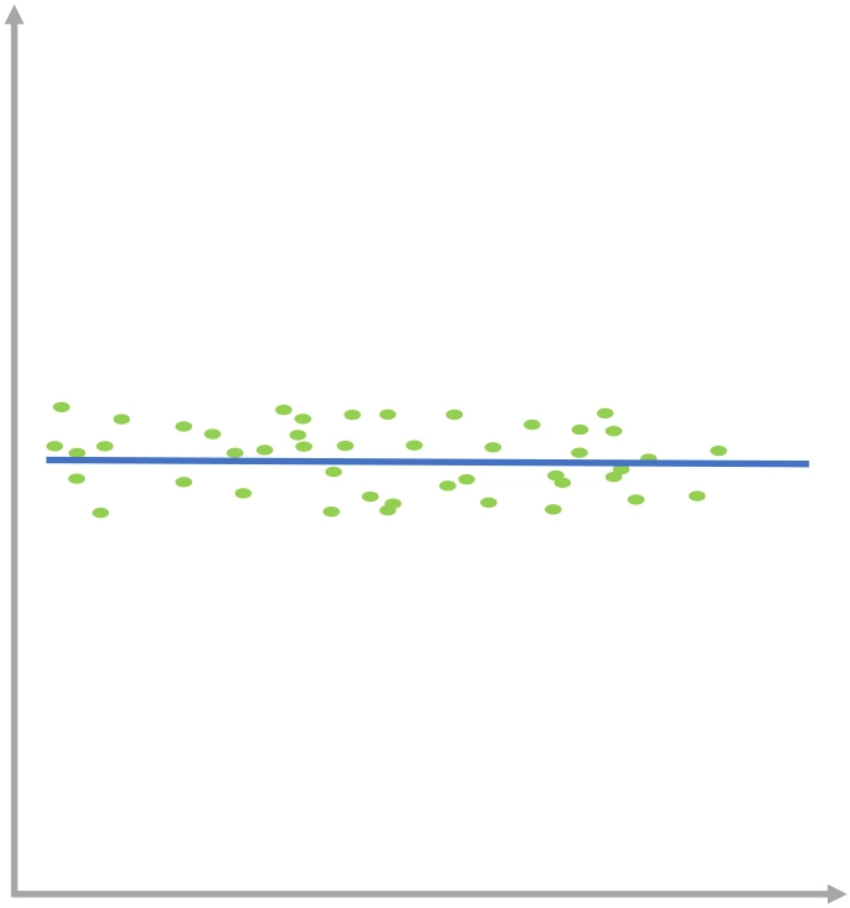
当税率开始增加时，税收一开始也会随之增加，但是当增加到一定程度，大家就交不起税了，重税导致企业开始倒闭和破产，这时候整体税收反而开始减小，到最终税率为 100% 时，其实企业会全部破产，税收反而是 0 了。



U 型曲线在很多场合也适用，例如员工工作时长和公司收入的关系（适度 996，不要 007），客户满意度与公司利润的关系（没有口碑的公司破产了，不加控制让客户全部高度满意的公司也都破产了）。

而**数据分析的艺术就在于通过数据分析和管理经验找到反 U 型最高点**，如果你能很好把握你所在公司的反 U 型曲线高点，你大概率是公司的管理层了。

还有一种情况就是所有的点全都分布在一条平行于 X 轴的横线两侧，就像下面这张图一样。其实这恰恰也表明了一个规律，就是这两个数据基本没有太大关系。也就是说不管 X 轴怎么变化，Y 轴的指标它就是我行我素，在这种情况下，如果 X 轴表示的是成本投入，你就不要再自欺欺人说将来还可以有收入增长了。



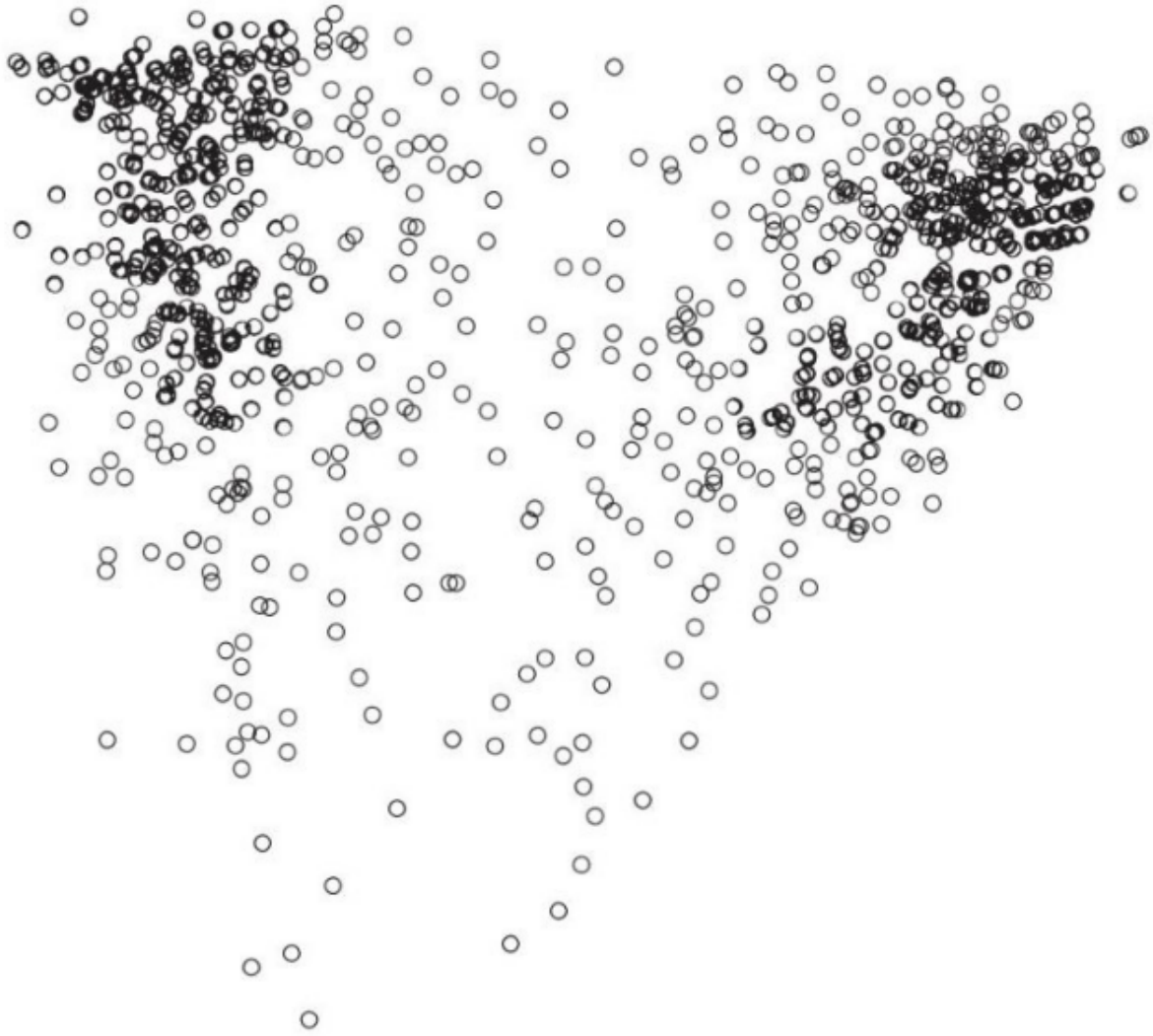
最后一种情况就是**散点图呈一个非常复杂的图形**。这个时候，我们不能轻易下结论，它需要我们根据所在的领域、行业的知识进行更细致的划分。

比如下面这张图就是挺著名的一幅散点图，这里面的数据分布就像一个心脏一样，心脏的两个左右上角的点聚集都比较密。但是你说它是什么趋势呢？貌似也看不出来。

这其实是当时美国大选进行民意调查时，选民主党和共产党的选民对 50 个不同调研问题的回答反馈总结出来的散点图。尽管我们看不出来这里面的趋势规律，但是我们还是能明显看到共和党和民主党选民的不同。

将来给你在算法部分讲聚类算法的时候，会再去给你深入讲这种情况怎么去划分。它不是用简单的线性、指数、二次项式等等趋势来表明的，它背后规律需要更复杂的计算机算法

才能够揭示出来。



散点图的易错点

看了这些，你会不会感觉散点图好像真挺万能的。但是你要注意，过度依赖散点图会造成你的判断错误。

最常见的三个错误就是趋势误判、得克萨斯神枪手谬误和幸存者偏差，接下来就是我们的“排雷时间”。

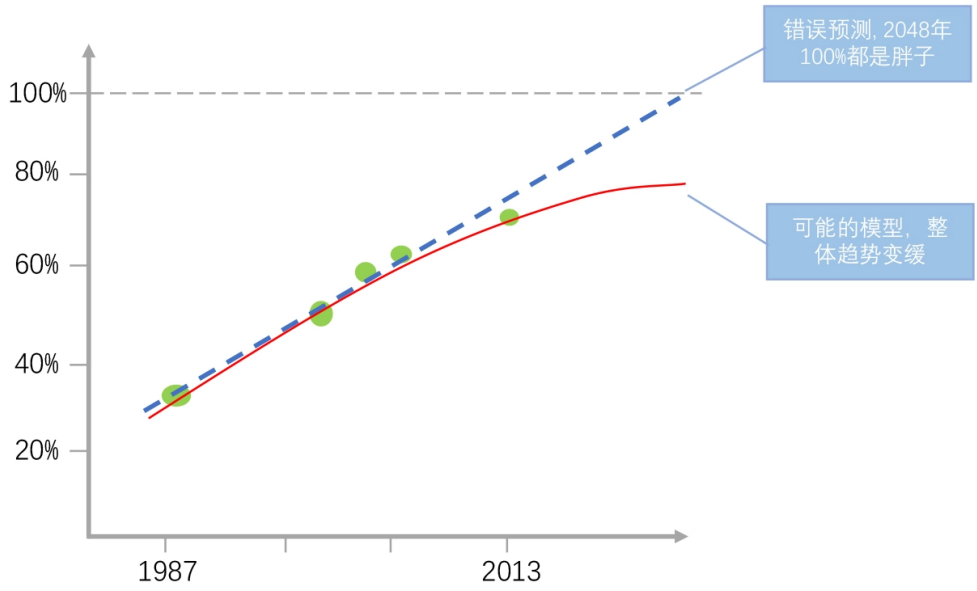
趋势误判

趋势误判是指在你看到了一个数据之后，因为数据整体还不够完整，你错误判断了这个数据的未来增长趋势。

这里我给你举个例子，曾经一位约翰斯·霍普金斯大学学者，给出了一个美国人体重增长趋势预测的散点图。这个散点图表示，在最近几十年内，美国人的超重现象越来越普遍。20 世纪 70 年代初，体重数超重的美国人不足半数；到 90 年代初，超重人数接近 60%；到 2008 年，几乎有 3/4 的美国人都超重了。

这个学者进行线性回归，其分析结果大致为：到 2048 年，这条线会达到 100%。因此，这个学者在论文中断言，如果这种趋势继续下去，到 2048 年，所有美国人都会超重，也就是 2048 年在美国每个人都会是胖子。这个论文受到了大肆吹捧，各种媒体都争相报道，连中国媒体都进行了转载，你如果感兴趣的可以点击这里，进一步了解一下。

这明显是个错误的结论。因为最终肥胖人群的增长趋势不是线性的，你现在看到的增长点似乎是线性，但其实它会是一个像抛物线一样的数据趋势（如下图）。



你看看，对于全球这么著名大学的学者来说，如果错误使用了散点图，都会得出错误的结论。所以当你在拿到一个散点图，要去判断它是哪种数据趋势的时候，一定要看最终的数字偏差和实际情况，才能做出准确判断。

趋势线这个画龙点睛的一笔，不是那么容易给出的。通过散点图最终判断是什么模型是非常不容易，很多时候需要非常多的数据和复杂的模型，这也最终导致了人工智能算法的出

现，当然这是后话了。现在你只要记住，**没有正确的数据验证，千万不要轻易下结论**，要不你也会和这个学者一样，闹一个大笑话。

得克萨斯神枪手谬误

在说这个谬误之前，先给你分享一个故事。

当年在美国西部得克萨斯州发现一个神枪手，他经常在各地的民居的墙上练习射击，几乎他所有的弹孔都集中在十环左右这个中心的区域。他已经成为了神话，人们一直在寻找他。

但是当人们真的找到了这个神枪手后，发现他自己打枪其实一点都不准，也不敢跟其他人去决斗。那他墙上的这些靶子和子弹点是怎么形成的呢？后来人们才发现，原来他是先朝墙上开很多枪，然后在弹孔最密集的地方画上了十环的靶子，再把散布在其它地方的弹孔用原来的泥土补起来。这样看上去，他每个地方打的靶子都很准确，因为先有弹孔，再有靶子。

在我们日常生活当中也很容易出现这种情况，当你看到一个数据散点报告的时候，你一定要看清**背后所蕴含的实际数据是不是涵盖了所有的数据，还是只给你看了最有这种数据规律的数据。**

前者就像先有靶子来瞄准再去射击，后者就好比先射击完最后再画上靶子，这样结果会完全不同。依据数据决策很重要，但是也不要被数据给骗了。

幸存者偏差

幸存者偏差这点我们在前面提到过，这里我再基于散点图给你强调一下。

我们小的时候，邻居家的小孩永远比我们要更厉害，其实孩子都是差不多的，只不过最后我们看到的是邻居家小孩当中的那些优胜者；自古红颜多薄命，也是因为我们只把目光放在了少数的红颜身上；天妒英才也是因为我们没有过多关注普通人究竟寿命几何。那对应到我们画散点图上也是一样的，**你在分析散点图的时候看到了规律，还要了解最终这个规律形成的原因和背后的场景，不要简单通过一个图表就得出你的结论。**

在工作和生活当中，每天都会有各种各样的事发生，如果我们只是关注事情本身，而没有看到背后的规律，那么我们就会像没有趋势线的散点图一样，都是零散的点，抓不住背后隐藏的那根线，感觉每天都是忙忙碌碌，但其实自己碌碌无为。

画这根线就是要找到你生命里的规律，在数据科学里我们叫做算法，生活当中我们叫做哲学。小小预告一下，在下一章算法里，我会和你继续深入探讨数据和客观世界背后的规律。

小结

好了，今天的这节课到这里也就接近尾声了，最后我们来总结一下。今天给你讲的是万图之王——散点图。小到我们自己投资和回报，大到整个宇宙里面星体之间的分布，我们都可以通过散点图来找到背后隐藏的规律。

要做一个正确的散点图，我们需要注意三个要点：

确定两个变量坐标轴；

坐标轴的起始值和颗粒度要合适；

要找到合适的趋势线和趋势模型进行描述。

紧接着我给你分享了几个在使用散点图时，容易犯的错误：

利用散点图做深入数据分析的时候不要轻易下结论（身高和体重在生长期是成正比的，你成年了自然也就不会再是正比了）；

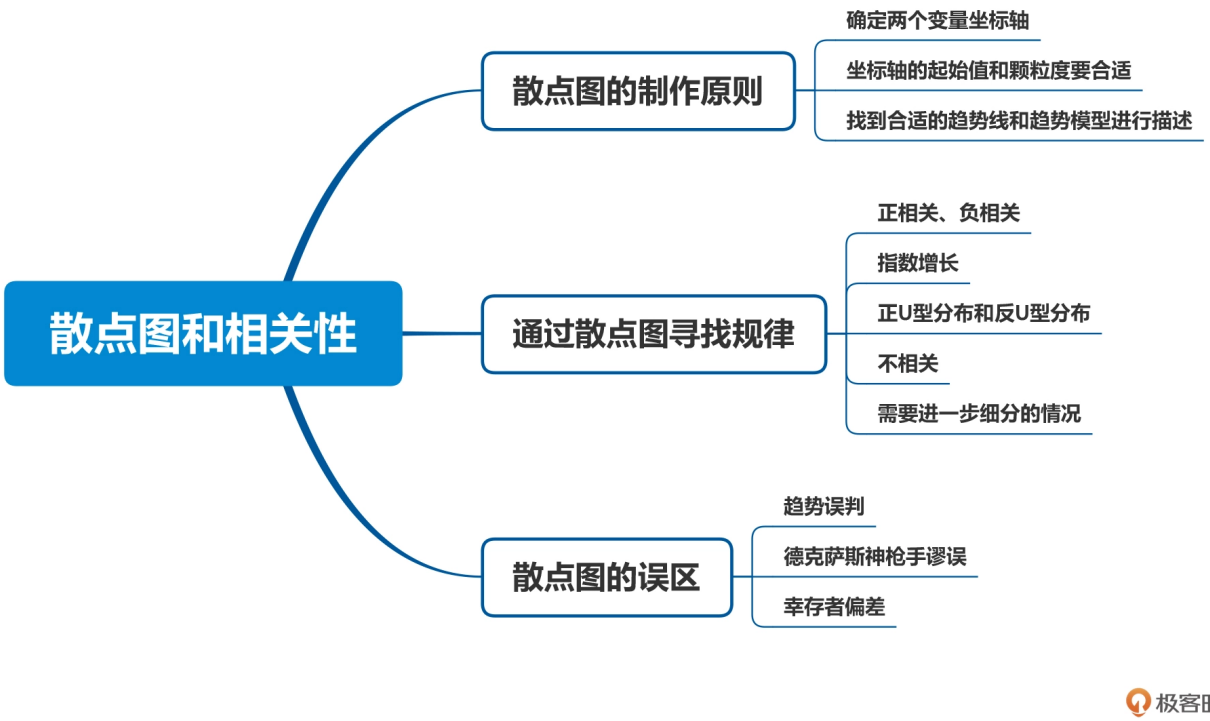
不能由现有结果给出趋势判断（做事情无论成功了还是失败了，都不要用上天的安排来麻痹自己），还需要了解规律形成的最终原因和背后的场景。

不要用片面的数据来证明你的规律（不要片面看问题，别人家的孩子真的不比我们好多少）；

你不妨试试用最近 48 个月投资股票和基金的累计回报来画出一个散点图。你来看看自己的投资的散点图到底是正相关、负相关、无相关还是指数增长呢？结合前面学习的大数定律，这个散点图也就会告诉我们处于投资经验的哪个阶段。

我们要根据实际自己的业务领域知识以及后面讲到的算法模型找到接近事实的最佳解，这样才能够帮助你去预测这个世界，而不要错误利用模型最后导致我们出现错误的决策。

数据给你一双看透本质的眼睛，希望你可以通过散点图加上准确的数据模型和业务知识，真的让数据帮助到你的日常生活和工作决策。数据知识学无止境，让我们一起持续学习，一起共勉。

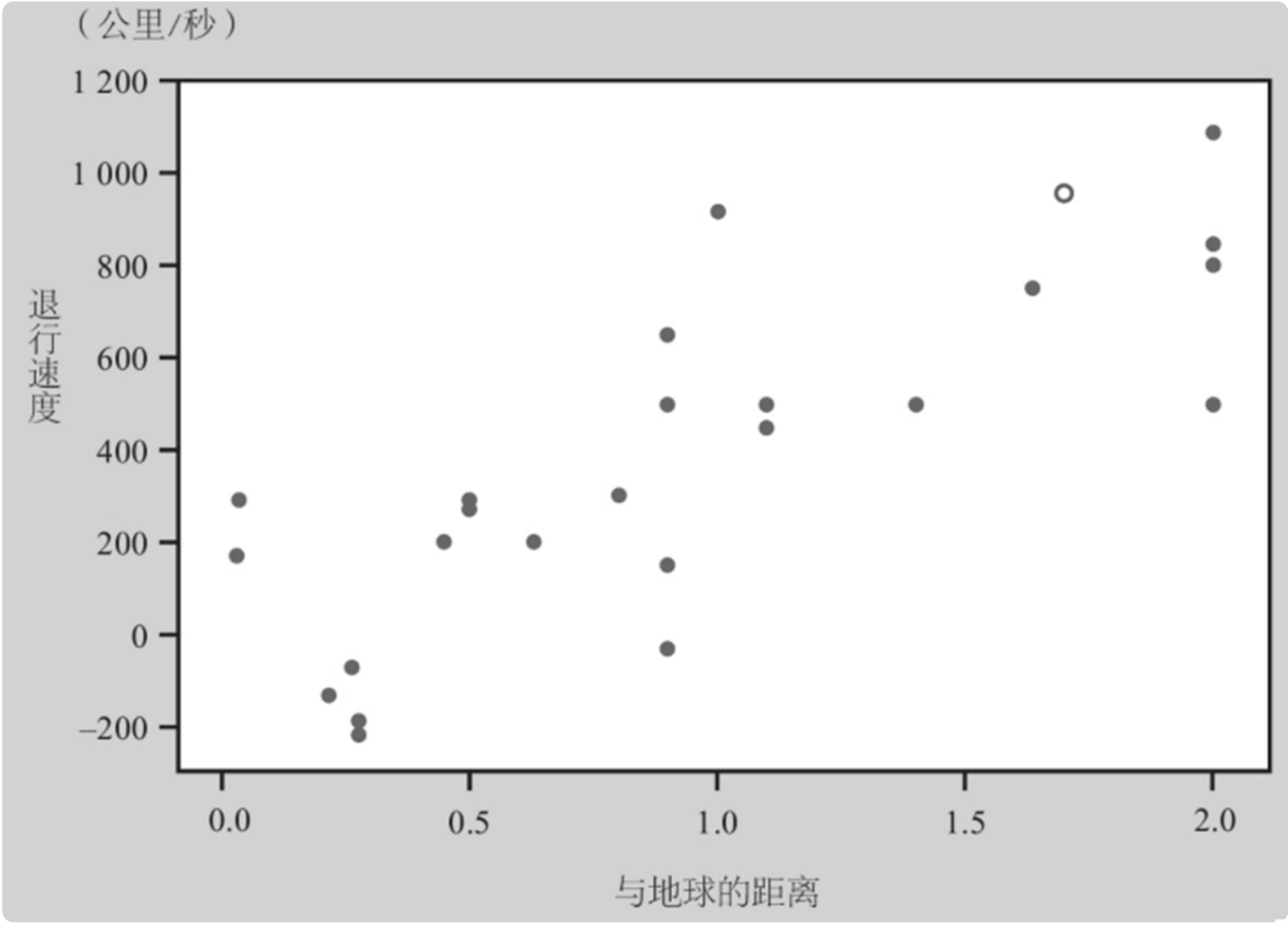


课后思考

你过去经验里，还有哪些是利用散点图发现的规律？哪些是错误利用散点图的教训？你生活中最常见的散点图是什么？欢迎分享给大家，我们一起学习。

附录 - 哈勃定律

哈勃（对，就是以他的名字命名哈勃天文望远镜那位） 也是用散点图展示了星系的退行速度和他们离地球距离之间的关系。下图里横轴是这些星系与地球的距离，纵轴是这些星系的退行速度。



这个散点图里图上每个点都代表了一个星系，通过这个图哈勃发现，与地球距离越远的这些星系退行速度越大，也成就了哈勃在天文学里找到的最重要的天文规律之一——哈勃定律。哈勃定律认为，星系可见的退行速度与它们和地球的距离成正比。总结出来的公式就是：退行速度 =H0 × 距离，其中 H0 就是哈勃常数。这个定律是支持大爆炸理论和解释宇宙膨胀思想的一个关键证据，也让哈勃成为近代最著名的天文学家之一。

分享给需要的人，Ta订阅后你可得 20 元现金奖励

赞 0 提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 06 | 数据分布：房子应该是买贵的还是买便宜的？

精选留言 (3)

智仔

2021-08-11

写留言

由數據看趨勢然後再找出規律……但問題是要有多少數據才能產生足夠信心的規律呢？感覺從數學問題要變成哲學問題…最後再變成神學思考了

展开

芳菲

2021-08-11

疫情和人员流动率的关系；
地震强度与持续时间的关系；

那时刻

2021-08-11

老师文中提到：散点图为了能够明确展示数据之间的趋势，我们的 Y 轴必须要从零开始。如果Y轴的值都是100以上，也必须从零开始么？

展开

1